Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications

Anna Watson
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT
+44 (0)171 419 3688
A.Watson@cs.ucl.ac.uk

M. Angela Sasse
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT
+44 (0)171 380 7212
A.Sasse@cs.ucl.ac.uk

1. ABSTRACT

There is currently much discussion of Quality of Service (QoS) measurements at the network level of real-time multimedia services, but it is the *subjective* quality perceived by the user that will determine whether these applications are adopted. This paper argues that ITU-recommended methods for subjective quality assessment of speech and video are not suitable for assessing the quality of many newer services and applications. We present an outline of what we believe to be a more suitable testing methodology, which acknowledges the multi-dimensional nature of perceived audio and video quality.

1.1 Keywords

Speech quality, video quality, subjective measurement.

2. INTRODUCTION

The number of real-time multimedia applications over packet networks has been increasing steadily, and with it the need to measure and assess the quality of multimedia delivered in this manner. There has been a surge in literature addressing Quality of Service (QoS) issues, but the emphasis has been on the quality of service at the *network* level, rather than from the end-user's point of view. Since it is the end-user who will determine whether a service or application is a success, it is vital to carry out *subjective* assessment of the multimedia quality delivered through these. There is an implicit assumption in parts of the networking community that QoS issues will eventually be resolved through implementing forms of bandwidth reservation (e.g. RSVP[1]) or increase (e.g. [2]), but as others recognize (e.g. [3]) there will also be consumer demand for lower quality at lower cost. Thus, it is important to establish the subjective quality boundaries for different real-time multimedia applications and the tasks they are used for. Designers of services and applications not only need to know *optimal* conditions for successful task completion, but the *minimum* quality required for a particular task, and the *maximum* point beyond which increased quality has no benefit for the user.

Before overall quality requirements can be tackled, it is necessary to investigate the perceptual influence of individual factors. The subjective impact of audio variables such as packet loss, delay, echo, background noise etc. needs to be considered. With respect to video transmission, available bandwidth and processing power can constrain the quality of the images that can be sent and received, and packet loss and delay can cause 'blocking' of the image and an irregular update rate. In addition, the subjective effects of network characteristics for some networks are more critical than for others. For example, packet loss over IP networks can cause severe damage to speech intelligibility, since audio packets often contain 40 or 80 msecs of speech information, matching the duration of the critical unit of speech comprehension, the phoneme. Although various methods of repairing packet loss in the audio stream have been investigated [3], overall perceived speech *quality* is not necessarily improved alongside an increase in intelligibility [4], illustrating the complexity of subjective quality measurement.

In this paper we present a critical review of existing methods of measuring subjective speech and video quality, before considering in more detail precisely what quality is, and how it should be measured in the context of real-time multimedia services and applications.

3. MEASURING PERCEIVED QUALITY

The most widely used methods for measuring the subjective quality of speech and video images have been standardized and recommended by the International Telecommunications Union (ITU). We consider these recommended methods now.

Quality of the speech/ connection	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

(a) Listening-quality scale

Did you or your partner have any difficulty in talking or hearing over the connection?

Yes	1
No	0

(c) Conversation difficulty scale

Effort required to understand the meaning of sentences	*Score*
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

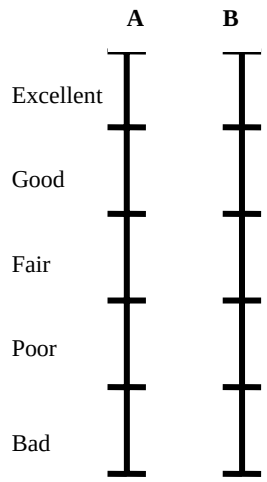
(b) Listening-effort scale

Image quality	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

(d) Image quality scale

Image impairment	*Score*
Imperceptible	5
Perceptible, but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

(e) Image impairment scale



(f) Double stimulus continuous quality scale

Figure 1: ITU recommended speech and image quality measurement scales

3.1 ITU Recommended Scales

ITU-T and ITU-R recommendations address subjective assessment of speech transmission over telephone networks and image quality over television systems, respectively. A series of ITU-T recommendations also address the subjective assessment of *multimedia* applications. The recommended scales are briefly presented below.

3.1.1 Speech Quality Scales

For the assessment of speech quality, the recommended rating scale for both listening-only and conversation tests is a 5-point category scale commonly known as the quality scale [5]. Listening-only tests can also be assessed via the listening effort scale. In conversation tests, a binary difficulty scale follows the (connection) quality scale. These scales are shown in Figure 1(a-c).

3.1.2 Image Quality Scales

For the assessment of image quality, single stimulus methods are rated using the quality scale or impairment scale, and comparisons to reference conditions are made using the double-stimulus continuous quality scale (DSCQS) or the double stimulus impairment scale [6]. These scales are shown in Figure 1 (d-f).

3.1.3 Audiovisual Quality Scales

Methods for the assessment of audiovisual communications are presented in [9]. The overall methodology is based on conversation opinion tests. The 5-point quality scale is recommended for assessing the video quality, the audio quality and the overall audiovisual quality. A 5-point 'effort needed to interrupt' scale can also be used.

We shall now consider the utility of these scales with respect to speech and video in real-time multimedia communication (MMC).

3.2 MMC Speech

Criticism of the recommended scales with respect to MMC speech falls into 3 main areas:

vocabulary of the scale labels;

length of the recommended test material;

conversation difficulty scale.

MMC speech is (in the main) narrowband and subject to a range of network and environmental degradations. Given these facts, the labels on the listening quality scale (i.e. Excellent, Good, Fair, Poor and Bad) seem inappropriate. Even with training, it is likely that responses will be concentrated at the lower end of the scale, which has been borne out in both experimental and field studies [7]. With respect to the category labels on the listening effort scale, it is even easier to see how a bias towards the lower end of the scale might occur.

The variable network conditions that affect some real-time services mean that speech quality can change rapidly and unpredictably. In listening-quality tests the recommended test material is short in duration – 10 seconds at most. This length of time does not afford the opportunity to experience the unpredictability of some

networks or, if loss rates are low, the full potential of the resulting impairment.

Finally, the binary difficulty scale is patently unsuited for the assessment of MMC conversations, since even a small amount of packet loss is likely to cause difficulty in hearing or talking, even if short-lived.

3.3 MMC Video

As with MMC speech, criticism of the recommended scales with respect to MMC video assessment falls into 3 main areas:

vocabulary of the scale labels;

duration of the test material;

artificiality of assessing video without audio.

The ITU-R recommendations are concerned with establishing the subjective performance of *television systems*. This means that in terms of color, brightness, contrast, frame rate etc., the quality component under investigation is assumed to be already of a high standard, which is simply not the case for MMC video. Like MMC speech, MMC video is characterized by a large variety and range of impairments, which can change rapidly. This trait means that the single- and double-stimulus impairment tests are not suitable, since, as is reflected in the terminology of the scale (*imperceptible/perceptible*), they have been designed to determine whether individual *small* impairments are detectable.

With respect to use of the quality scale, the same criticism can be leveled as to its use with MMC speech: the vocabulary is unsuitable, and therefore we can expect responses to be biased towards the bottom of the scale. Use of the DSCQS at least permits scoring between the categories (the subject places a mark anywhere on the rating line, which is then translated into a score), but it is still the case that subjects shy away from using the high-end of the scale, and will often place ratings on the boundary of the 'good' and 'excellent' ratings [8].

The quality tests typically require the viewer to watch short sequences of approximately 10 seconds in duration, and then rate this material. It is not clear that a 10-second video sequence is long enough to experience the types of degradations common to MMC video. This problem will be discussed further in section 5.3.

In addition, the quality judgements are intended to be made entirely on the basis of the picture quality. It should be queried whether it makes sense to assess MMC video on its own (i.e. without audio) since it would be true to say that the video image in MMC is not the focus of attention in the same way that the picture is when we watch television. We believe that the utility of the low frame rate video currently used in MMC arises mainly when it is used in conjunction with audio (and perhaps shared workspace), and so it is only in real task environments that it makes sense to evaluate the subjective quality of the video. It would be highly unusual, if not inconceivable, for users to be using low-frame rate video as the sole means of communication across networks at present. For this reason, the audiovisual quality recommendations should be better suited to assessing MMC video. However, since it is the 5-point scales that are recommended again, the criticisms raised above remain valid. 'One-off' quality ratings gathered at the end of an audiovisual session also do not capture the changing perceptions users may have during communication across a packet network with varying conditions (see section 5.3).

We have argued that the assessment methodologies recommended by the ITU are not suitable for subjective quality assessment of MMC over packet networks. In particular we have argued that the 5-point quality scales are not viable due to their vocabulary. But there is a yet more serious issue at hand – how legitimate are the 5-point scales to begin with?

3.4 The Nature of the International Interval Scale

The 5-point quality scale is easy to administer and score, and its recommendation by bodies such as the ITU has meant that its use has been accepted without question by many researchers. There are a growing number of researchers, however, who question whether such trust in this scale is warranted. Investigations have focused mainly on whether the quality scale is actually an interval scale, as represented by the labels on the categories. If the intervals on the scale are not equal in size, then it is doubtful whether the use of parametric statistics on the data gathered from quality assessments is strictly legitimate, since this would require a normal distribution [10]. Investigations have also been carried out to validate the ITU assumption that the scale labels have been adequately translated into different languages, such that the scale is 'equal' in different countries, so that quality results can be generalised across the world.

3.4.1 Internationally Interval, or Internationally Ordinal?

Investigations of the interval nature of the rating scales have generally been carried out using the graphic scaling method. Subjects are presented with a vertical line with the words "Worst Imaginable" at the bottom, and "Best Imaginable" at the top. On this line, they are required to place a mark where they feel a certain qualitative term would fit. By measuring the distance of the marks from the bottom of the scale, the means and standard deviations for each term can be calculated. Using this method, Narita [11] found that the Japanese ITU labels conform well to the model of an interval scale, although not perfectly. Whilst this is good news for Japanese speakers, it is a different story for English, Dutch, Swedish and Italian speakers.

Jones & McManus [10] used the same method to investigate whether the intervals represented by the quality scale labels are equal i.e. that the distance between 'Good' and 'Fair' is equal to the distance between 'Poor' and 'Bad'. They found that the scale terms were spaced almost as a 4-point, 3-interval scale as opposed to the 5-point, 4-interval scale they are supposed to represent i.e. the ITU terms constitute an ordinal rather than an interval scale. 'Bad' and 'Poor' were found to be perceived as very similar in meaning, whilst the perceptual distance to 'Fair' was comparatively great. Since research in psychology has established that subjects tend to avoid the end points of scales, they question the usefulness of what appears essentially to be a "3-point, 2-interval scale".

Jones & McManus also carried out their study in Italy. The Italian ranking of the ITU terms produced a scale that has no mid-point. In the ranking of other terms, it is interesting to note that a supposed 'universal' word such as 'OK' appears to mean different things to different nations: the Americans positioned 'OK' around the centre of the scale, as roughly equivalent to 'Fair', whereas the Italians seemed to equate 'OK' with 'Good'.

Other researchers have found similar results. Virtanen et al. [12] found that there was a flattened lower end (i.e. the Swedish terms equivalent to 'Bad' and 'Poor' were perceived as very similar), and there was a large gap between 'Poor' and 'Fair' such that 'Fair' was actually above the midpoint of the scale. Teunissen [13] investigated Dutch terms and found once more that the ITU terms do not divide the scale into equal intervals.

3.5 Summary

The ITU-recommended quality scale is not the international interval scale it is purported to be. But the quality scale is also not internationally *ordinal*, since the positional rankings of the qualitative terms in different languages are not equal. However, there is another, more complex issue at hand, and that is the overall concept of quality: the 5-point quality scale treats quality as a single measurable dimension, despite much evidence to the contrary.

4. WHAT IS MULTIMEDIA QUALITY?

Virtanen et al. [12] demonstrated that quality is not a "single monotone dimension" - or at least the terms used to describe it are not. They investigated the semantic groups that qualitative terms fall into, and determined that there are at least 4 types of quality scaling situations: qualitative/hedonic judgement, positioning in relation to a reference, emotional/communicative expression and 'people as judges'. The existence of so many quality categories highlights the fact that many different variables can affect quality perception formation. What can we say about the variables that contribute to speech and video quality perception?

4.1 Speech Quality

Researchers from disciplines as diverse as hearing aid research and engineering have identified significant roles in speech quality for variables such as intelligibility, loudness, naturalness, listening effort, pleasantness of tone etc. [e.g. 5,14]. However, as Preminger & Van Tasell [14] observe, "*Although a multidimensional view of speech quality has not been disputed, many researchers have taken a unidimensional approach to its investigation… When speech quality is treated as a unidimensional phenomenon, speech quality measurements are essentially judgements, and one or several of the individual quality dimensions may influence the listener's preference*." This approach does not allow us to determine which of the many factors that comprise quality carry most weight in perception formation.

Just as there is a unidimensional approach to measuring quality, within the networking community there is also a tendency to assume a unidimensional approach to improving quality: increasing bandwidth. For example, "*the notion of quality as a function of speech bandwidth will become more pervasive, and subjective testing will lead to better quantification of the quality-bandwidth function*" [2]. However, although increasing bandwidth would undoubtedly solve many quality issues, it should not be treated as a panacea. It may well be the case that many quality issues can be settled without resorting to increasing bandwidth, and since bandwidth is a valuable resource, exploring these possibilities is important, for both the HCI and networking communities [7, 4, 3].

4.2 Video Quality

Subjective opinion of video quality is also formed through the influence of many different factors. Gili et al. [15] identified seven key variables to be color, brightness, background stability, speed in image reassembling, outline definition, 'dirty window', and the mosaic/blocking effect.

However, for MMC video it is perhaps more important to investigate the *interaction* between speech and video when considering the quality determinants (see section 3.3). The overall benefits to combined audio and video are far greater than when taken individually and summed. The importance of the task being undertaken to the quality perception of video and speech should also not be underestimated [7].

4.3 Summary

Speech and video qualities are multidimensional phenomena. We must develop a means of identifying these different dimensions. Once the components have been identified, it then becomes necessary to determine their relative impact on overall subjective quality for different tasks. This process requires a new quality scaling method. Preliminary steps towards this goal are presented in the following section.

5. NEW APPROACHES TO QUALITY MEASUREMENT

Extracting and measuring the quality components of MMC speech and video is being undertaken at UCL using a number of methods. Our three main goals are to

identify suitable vocabulary to describe subjective quality;

identify the key quality dimensions;

employ this knowledge in developing a new subjective quality measurement method.

5.1 Generating Suitable Vocabulary

We are aware that our own 'expert' vocabulary battery for describing MMC speech and video quality is limited, and we are concerned that our 'technical' descriptors do not match the terms and concepts that the general population would use to describe their perceptions. The use of open-ended questions on quality, and encouraging participants in experiments and field trials to comment about the speech and video they have experienced, has allowed us to begin building a database of commonly used descriptive terms. This database serves two purposes. Firstly, we propose to have subjects rank the collected terms using the graphic scaling method in order to investigate whether we can develop more meaningful labeled rating scales for MMC speech and video. Secondly, the terms can act as a confirmation to the data gathered from focus groups and submitted to grounded theory analysis, as discussed in the next section.

5.2 Identifying the Quality Dimensions

We have begun to identify the different dimensions of quality, and which vocabulary terms, gathered by the methods described above, relate to these dimensions. For example, a key quality dimension for MMC speech has been identified as "choppiness", where associated quality terms are 'broken', 'cut up' and 'irregular'.

5.3 Investigating New Scales

Once we are confident that we have identified the key quality dimensions and the related vocabulary, we will require a means of assessing perceived quality along the dimension in question. We have investigated the use of an unlabelled continuous rating scale, in both controlled experimental studies and in field trials, and feel that this method would be suitable for rating along specified dimensions.

In [4], 24 subjects rated the quality of speech passages on a 200 mm unlabelled continuous scale, with a plus and minus sign at opposite ends of the scale to indicate polarity. We found that the quality rating results gathered from this have been remarkably consistent, considering that the subjects set their own criteria (see figure 2). We have also observed that using an unlabelled scale reduces the tendency of subjects to avoid the end points of the scale. In addition to the speech experiment, this unlabelled scale has been used to effect in a video quality experiment and a distance learning field trial. However, one major concern in both our controlled experiments and field trials has been the length of the 'test' material i.e. the duration of the session that is being assessed.

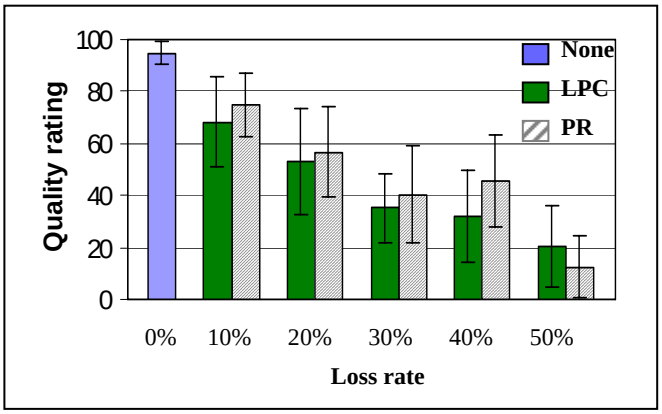


Figure 2: Listening quality results using an unlabelled continuous quality scale, where two different packet loss repair schemes, LPC and PR, are compared.

In sections 3.2 and 3.3, it was observed that the recommended test material length is often unsuitable for the assessment of new communication technologies. For example, investigators of subjective quality assessments for ATM video impairments found that the ITU-R recommended test sequence length (approx. 10 seconds) is not long enough to capture the range of impairments that are typically found in ATM video [8]. If the test sequences are extended, the recommended DSCQS method cannot then be employed, since the load on memory becomes too great [16].

The load on memory that arises from longer assessment periods is a problem that should not be taken lightly. A concern that arose from our field trials was that the quality ratings given at the end of an hour-long conference were cumulative i.e. it was not possible to know what parts of the conference had the greatest influence in forming the judgement [7]. Studies of video quality assessment have pointed to the likelihood of recency effects playing a role. Seferidis et al. [17] reported on what they termed the "forgiveness effect " in which observers 'forgive' impaired video when it is followed by a substantial period of unimpaired video. It has furthermore been discovered that when good quality video precedes poor-quality, the rating will be awarded on the basis of the poor-quality section, thus linking this phenomenon to the recency effect of memory [8]. It seems likely that if observers are asked to give a single quality rating at the end of a video segment, they will be significantly influenced by what they saw in the last part of the segment. It is likely that this is the case for speech quality rating also.

As test stimuli become longer, another confounding issue to be aware of is increasing interest (or boredom!) with respect to the test material. Aldridge et al. [8] reported that some observers were "distracted" in their task of quality assessment by the content of the video sequence, and we have observed this effect too in our own MMC studies [4]. Moreover, Wilson et al. [18] found that an increase in task difficulty may have the effect of decreasing the subjective image quality, a finding consistent with cognitive dissonance theory.

The presence of confounding issues such as these in quality judgements gives weight to an argument for a more dynamic, instant means of measurement. de Ridder & Hamberg [16] provided observers with a slider mechanism labeled with the Dutch quality scale terms. The observers manipulated this slider as they watched video sequences, and the results showed that they were able to monitor video quality variations as they occurred.

We are currently investigating the utility of a software version of a dynamic slider., QUASS (QUality ASsessment Slider). The scale used is the unlabelled continuous scale discussed above. The slider bar on the scale is operated by mouse, and measurements of the slider's position are taken every second, allowing us to match subjective results with known objective conditions. QUASS has a twofold functionality. In the first scenario subjects use it to continuously rate perceived quality along a specified dimension, allowing us to relate perceived quality to a precise instant of the test material. In the second scenario, the subject is able to *control* the quality dimension under investigation via the slider.

Our initial study with QUASS has been a laboratory audio-only study. Final analyses on the data have yet to be carried out, but observations of the tool in use are encouraging. We are currently implementing QUASS for use in a range of MMC project tasks over the Mbone. We hope this approach will enable us to begin establishing subjective quality requirements for different types of conferences, since we will be able to compare our subjective results with objective data such as RTCP reception statistics [19].

6. SUMMARY AND CONCLUSIONS

Researchers, network providers and application developers have a requirement to understand and measure the perceived quality of

real-time multimedia communication from the end user's point of view. We have summarized here a growing body of evidence which indicates that results obtained from existing rating scales - which were developed to assess quality for very different types of networks and applications – may be imprecise at best, and thoroughly misleading at worst. Although these scales have more than proved their worth in many communication arenas, they should not be used to assess subjective quality required by multimedia applications developed today, or used to infer bandwidth or other QoS requirements for network services. There is a necessity for reliable and valid methods to measure subjective speech and video qualities in the applications developed, and link them to the objective QoS factors that can be applied to network services.

In section 5 we outlined an approach to assessing audio and video quality which addresses this requirement. It acknowledges that there are multiple factors that influence users' perception of multimedia speech and video. On the basis of reported literature, and a number of field trials, experiments and focus groups, we have identified a set of dimensions that we believe determine users' perception of quality in a large number of tasks and situations. We propose a set of methods, which we have evolved in our empirical work, to measure user perception for each of those dimensions. Our aim is to be able to pinpoint actual *quantities* for the dimensions, i.e. establish the critical quality boundaries (minimum and maximum quality thresholds) for a particular dimension in the context of a particular task. Once a large set of empirical data has been collected, this approach would yield a taxonomy of quality boundaries for audio and video for a range of tasks. Applications developers and service providers could apply the taxonomy to infer objective QoS requirements for particular applications.

7. Acknowledgements

We gratefully acknowledge the contributions of Anna Bouch and Louise Clark from UCL Computer Science. Anna Watson is funded through an EPSRC CASE studentship with BT Labs.

8. References

[1] Zhang, L., Deering, S., Estrin, D., Shenker, S. and Zappala, D. RSVP: A new resource ReSerVation Protocol, IEEE Network Magazine, 1993, 7(5), 8-18.

[2] Jayant, N.S. High-quality coding of telephone speech and wideband audio. IEEE Communications Magazine, Jan. 1990, 10-20.

[3] Podolsky, M., Romer, C. and McCanne, S. Simulation of FEC-based error control for packet audio on the Internet. Proc. INFOCOM '98.

[4] Watson, A. and Sasse, M.A. Multimedia conferencing via multicast: determining the quality of service required by the end user. Proc. AVSPN '97.

[5] ITU-T P.800 Methods for subjective determination of transmission quality.

[6] ITU-R BT.500-7 Methodology for the subjective assessment of the quality of television pictures.

[7] Watson, A. and Sasse, M.A. Evaluating audio and video quality in low-cost multimedia conferencing systems, Interacting with Computers, 1996, 8 (3), 255-275.

[8] Aldridge, R., Davidoff, J., Ghanbari, M., Hands, D. and Pearson, D. Measurement of scene-dependent quality variations in digitally coded television pictures. IEE Proc.-Vis. Image Signal Process, 1995, 142(3), 149-154.

[9] ITU-T P.920 Interactive test methods for audiovisual communications.

[10] Jones, B.L. & McManus, P.R. Graphic scaling of qualitative terms. SMPTE Journal, November 1986, 1166-1171.

[11] Narita, N. Graphic scaling and validity of Japanese descriptive terms used in subjective-evaluation tests. SMPTE Journal, July 1993, 616-622.

[12] Virtanen, M.T., Gleiss, N. and Goldstein, M. On the use of evaluative category scales in telecommunications. Proc. Human Factors in Telecommunications '95.

[13] Teunissen, K. The validity of CCIR quality indicators along a graphical scale. SMPTE Journal, March 1996, 144-149.

[14] Preminger, J.E. and Van Tasell, D.J. Quantifying the relationship between speech quality and speech intelligibility. Journal of Speech and Hearing Research, 1995, 38, 714-725.

[15] Gili Manzanaro, J., Janez Escalada, L., Hernandez Lioreda, M., Szymanski, M. Subjective image quality assessment and prediction in digital videocommunications. COST 212 HUFIS Report, 1991.

[16] de Ridder, H. and Hamberg, R. Continuous assessment of image quality. SMPTE Journal, February 1997, 123-128.

[17] Seferidis, V., Ghanbari, M. and Pearson, D.E. Forgiveness effect in subjective assessment of packet video. Electronics Letters, 1992, 28(1), 2013-2014.

[18] Wilson, F., Wakeman, I. and Smith, W. Quality of service parameters for commercial application of videotelephony. Proc. Human Factors in Telecommunications '93.

[19] Schulzrinne, H., Casner, S. and Frederick, R. RTP: A transport protocol for real-time applications. IETF Audio/Video Transport Working Group, January 1996, RFC 1889.